УДК 004.056

М. С. Подгорный
Научный руководитель: канд. тех. наук, доц. К. А. Паршин
Уральский государственный университет путей сообщения, Екатеринбург

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ ИСПОЛЬЗОВАНИЯ МЕТОДОВ АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ НА ПРЕДПРИЯТИИ ЖЕЛЕЗНОДОРОЖНОГО ТРАНСПОРТА

Аннотация. Рассмотрены проблемы возможной утечки информации в сфере железнодорожного транспорта. Описаны особенности синтаксиса терминологии и описания объектов и субъектов на железнодорожном транспорте, приведены примеры сокращений должностей сотрудников. Рассмотрены возможные методы анализа текстовых данных на предприятии для каждой проблемы.

Ключевые слова: железнодорожный транспорт; информационная безопасность; утечка информации; лингвистический анализ; сентимент-анализ; синтаксис терминологии; байесовский наивный классификатор.

Железнодорожный транспорт является основной транспортной деятельностью в Российской Федерации. По отчетным данным вклад ОАО «РЖД» в ВВП РФ составляет 2,5 %, средний оборот денежных средств достигает практически двух триллионов рублей, а количество сотрудников — достигает 890 тысяч человек [1]. Сеть железнодорожных дорог протянута по всей территории Российской Федерации и включает 16 железных дорог.

Вышеописанные факты подтверждают актуальность сохранения состояния информационной безопасности на предприятиях железнодорожного транспорта. Ключевым моментом с точки зрения информационной безопасности для предприятия с численностью почти 900 тысяч человек является вопрос утечки информации ограниченного доступа. Ввиду развития социальных сетей как программно, так и количественно (в ежесуточной аудитории), а также других интернет-сервисов, как пример утечки информации ограниченного доступа стоит рассматривать и публикации определенных документов в общем доступе как «один для многих», содержащие «чувствительную» информацию для предприятия. Другими словами, с этими данными, которые опубликовал сотрудник предприятия, может ознакомиться определенная группа других людей (через публикацию в группе, «на стене» или через комментарий в определенной записи).

В качестве примера можно привести публикацию фотографий со служебной перепиской при столкновении двух пассажирских поездов на Московской

железной дороге, а также текстовых комментариев к ним на одном из известных общественных сайтов в сети Интернет. Публикация содержала информацию о служебной переписке поездного и станционного диспетчера, а также машинистов локомотивов [2]. Для предприятия данные текстовые записи несут как утечку информации ограниченного доступа, так и негативное представление компании в общем доступе. Данный частный случай подтверждает важность как мониторинга действий сотрудников в части взаимодействия с внешними сервисами, так и мониторинг самих сервисов.

Многие предприятия используют системы предотвращения утечек информации (Data Leak Prevention, DLP), однако не всегда запись может быть сделана из корпоративной сети. Другим важным фактом являются финансовые ресурсы предприятия и мировые экономические тенденции, что напрямую влияет на отказ использования дорогостоящих систем и поиск других альтернативных программных продуктов. Ключевым моментом, на котором стоит заострить внимание, являются аналитические методы, используемые в DLP-системах для работы с текстовыми документами и которые могут быть применены и в других прикладных программных средствах:

- поиск по регулярным выражениям;
- лингвистический анализ;
- байесовский алгоритм;
- морфологический анализ;
- предустановленные текстовые шаблоны;
- предустановленные тематические словари.

Для дальнейшего описания требуется раскрыть синтаксис сокращенных наименований объектов и субъектов, которые используются на железнодорожном транспорте. Для описания должностей, объектов инфраструктуры и прочих объектов используется специальная терминология, единая на всей территории страны [3]. В качестве примера, сокращение ДНЦ обозначает поездного участкового диспетчера, а НГ — главного инженера железной дороги.

Первой отличительной чертой железнодорожной терминологии является то, что объекты имеют определенную условную иерархичность при описании субъектов. Например, следующее описание должностей:

- Д — служба перевозок;
- ДС — начальник станции;
- ДСП — дежурный по станции;
- ДСПГ — дежурный по сортировочной горке;
- ДСПГО — оператор при дежурном по сортировочной горке.

Вторым признаком является именно синтаксический состав и порядок букв в сокращении железнодорожных объектов и субъектов. Согласные буквы

составляют 80 % во всех сокращениях, при этом гласные буквы расположены либо в начале, либо в конце краткого наименования объекта.

После анализа предметной области сформированы две возможные проблемы для предприятия — публикация в открытых источниках информации ограниченного доступа и негативная тональность данных публикаций. Как следствие, должны быть подобраны и аналитические методы работы с текстовыми данными непосредственно для решения этих проблем.

Для решения задачи определения тональности сообщений на первом этапе был сформирован классификатор на основе наивного байесовского алгоритма, а также основных методов сентимент анализа. В качестве обучающей выборки использовались корпуса открытых данных в виде коротких строковых сообщений из социальной сети Twitter, заранее классифицированные на две категории — негативные и позитивные отзывы [4]. В процессе построения классификатора был выполнен стандартный алгоритм «bag of words», выполнено удаление стоп-слов, входящих в русский словарь библиотеки *NLTK* языка программирования Python, а также выполнено приведение слов к начальной форме.

Для решения второй задачи поиска текстов, которые могут содержать ту или иную информацию ограниченного доступа выбран лингвистический анализ документа с заранее предопределенным текстовым словарем. Выбор данного метода сделан ввиду специфической терминологии, которая используется на железнодорожном транспорте. На первом этапе выполнения данной задачи планируется собрать словарь сокращенных наименований объектов и субъектов сферы железнодорожного транспорта для дальнейшего изучения данного вопроса [3].

Тестовая проверка будет выполнена на данных, полученных со стены пользователей из социальной сети «ВКонтакте», которые подходят по фильтру поиска город — Екатеринбург и работа — ОАО «РЖД». Данные получены через API данной социальной сети и содержат текстовые данные конкретных пользователей [5]. По результатам тестовой проверки и рассмотрения показателей классификаторов будет сделан вывод о возможности использования данных аналитических методов при работе с текстовыми данными.

Список литературы

1. Показатели основной деятельности [Электронный ресурс]. URL: http://ir.rzd.ru/static/public/ru? STRUCTURE_ID=63 (дата обращения: 10.10.2017).

2. Про столкновение электрички и поезда [Электронный ресурс]. URL: http://www.yaplakal.com/forum15/st/175/topic1579951.html (дата обращения: 11.04.2017).

3. Железнодорожный словарь [Электронный ресурс]. URL: http://rzd.me/inform-block/zhd-slovar/ (дата обращения: 20.09.2017).

4. *Рубцова Ю. В.* Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1 (109). С. 72–78.

5. Документация API социальной сети «ВКонтакте» [Электронный ресурс]. URL: https://vk.com/dev/manuals (дата обращения: 29.09.2017).

УДК 004.891

Е. С. Подоплелова
Научный руководитель: д-р тех. наук, профессор А. Н. Целых
Инженерно-технологическая академия
Южного федерального университета, Таганрог

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ. АЛГОРИТМ C4.5

Аннотация. В этой статье описывается алгоритм интеллектуального анализа данных C4.5. Описывается работа классификатора, его структура, пример. Представлен механизм построения дерева решений. Обосновывается актуальность и особенности рассматриваемого метода.

Ключевые слова: интеллектуальный анализ данных (data mining); экспертная система; алгоритм анализа данных; классификатор; дерево решений; атрибуты.

Интеллектуальный анализ данных (Data Mining) — это неотъемлемая часть любой экспертной системы. Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах. Технологии Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

Выбор конкретного алгоритма анализа данных остается за разработчиком на этапе проектирования системы. Существуют уже довольно распространенные, широко используемые варианты. Один из них — C4.5. Основан он на дереве решений. Приобрел популярность благодаря довольно понятному представлению и качественному механизму.

Классификатор. C4.5 строит классификатор в виде дерева решений (рис. 1). Для этого в C4.5 дан набор данных, представляющих вещи, которые уже классифицированы. Классификатор — это инструмент интеллектуального анализа